Wahrscheinlichkeitslehre

Hans-Peter Beck-Bornholdt, Hans-Hermann Dubben

Der Schein der Weisen - Irrtümer und Fehlurteile im täglichen Denken

Hoffmann und Campe, 2001, 207 Seiten, ISBN 3-455-09340-X

Irren ist menschlich und Irrtümer werden systematisch gemacht, vor allem in der Wissenschaft, Urteile brauchen Vorurteile. Die meisten medizinischen Studien über die Wirksamkeit neuer Medikamente oder neue Therapien sind falsch oder wertlos und nicht viel mehr als wenn man (mit weniger Kosten) gewürfelt hätte.

Mit solch provozierenden Thesen entführen die Autoren die Leser in die Welt der Vorhersagen und der Wahrscheinlichkeitsaussagen. Ein schwieriges Gebiet, denn solange Ursache und Wirkung direkt im Sinne „wenn eine Bedingung gegeben ist, tritt eine bestimmte Wirkung ein„ zusammenhängen, ist die Welt „noch in Ordnung„. (wenn es regnet, ist die Straße nass)

In vielen Lebensbereichen ist man aber auf Wahrscheinlichkeitsaussagen angewiesen. Und da macht der „gesunde Lebensverstand„ nicht so einfach mit. Die Aussage: „Auf Abendrot folgt am morgen Regen„ ist nicht 100%ig sicher, sondern tritt nur mit einer bestimmten Wahrscheinlichkeit ein. Auch ohne Abendrot kann es tags drauf regnen, mit Abendrot kann der morgendliche Regen ausbleiben. Wahrscheinlichkeitsaussagen sind daher immer mit einem bestimmten Fehler verbunden. Die Fachwelt spricht hier von „falsch positivem„ Fehler, z.B. wenn bei einem gesunden Patienten eine Krankheit diagnostiziert wird (Anm. „positiv„ bezieht sich nicht auf die Krankheit, sondern auf die Vorhersage) oder einem „falsch negativen Fehler„, wenn ein kranker Patient als gesund aufgefasst wird.

Natürlich sollten die Fehler möglichst gering sein. Hier hat sich als internationale Konvention herausgebildet, dass die so genannte „Irrtumswahrscheinlichkeit„ d.h. dass zufällig ein falsch positives Ergebnis herauskommt unter 5% liegt.(sog. p-Wert < 0,05)
Dann wird ein Ergebnis als „statistisch signifikant„ angesehen. Tausende von Studien in allen wissenschaftlichen Bereichen und v.a. in der Medizin werden nach diesem Schema durchgeführt nach dem Motto: wenn $p < 5\%$, dann ist eine neue Erkenntnis geboren.

Halt rufen da die Autoren: „Irrtum, Irrtum, Irrtum„. Da würden wichtige Dinge vollkommen vergessen. Eine wesentliche Frage sei doch, wie sehr oder wie stark in einer Studie überhaupt ein Effekt nachgewiesen werden könne, sofern eine bestimmte Bedingung vorliegt – also das, was die Statistiker die Macht, Kraft oder „power„ einer Studie nennen. Obwohl so wichtig, werde dieser Wert "power" meist nicht bestimmt oder angegeben. Dies bedeutet, dass eine Studie nach dem Kriterium $p<5\%$ „signifikant„ sein kann, aber trotzdem nur eine geringe „power„, d.h. eine geringe Aussagekraft hat. Damit hängt auch das Problem zusammen, dass die (falsche) Zuordnung von Personen zu bestimmten Gruppen in epidemiologischen Studien die „signifikanten„ Ergebnisse regelrecht „verwässern kann oder dazu führt, dass systematisch bestimmte Wirkungszusammenhänge nicht entdeckt werden können. Will man z.B. erforschen, ob Lösemittel ZNA-Schäden verursachen und man bildet zwei Gruppen, nämlich eine aus unbelasteten und eine aus Personen, die gegenüber Lösemitteln exponiert waren, rekrutiert aber die

Personen der ersten Gruppe der angeblich Unbelasteten aus dem Kreis von z.B. KfZ-Mechanikern, dann wird sich mit Sicherheit im Vergleich der beiden Gruppen kein Effekt zeigen, der mit Lösemitteln in Zusammenhang gebracht werden könnte. Nehmen überdies auch nur wenige Personen an der Studie teil, verschwinden alle Unterschiede.

Doch es kommt noch toller. Ohne dass bestimmte Voraussetzung in eine Untersuchung hineingesteckt werden oder vorliegen, kann das Ergebnis eine ganz verschiedene Wertigkeit oder Bedeutung haben. Ein Beispiel: Eine Regenvorhersagemaschine kann eine gute Signifikanz aufweisen. In nur 5% aller Fälle arbeitet sie „falsch positiv„ und sagt Regen voraus, und es bleibt trocken. Doch und da ist der Haken, die Vorhersage mag eine interessante Sache in Gebieten sein, in denen es *typischerweise* an der Hälfte der Tage im Jahr regnet oder nicht. Überträgt man das jedoch auf trockene Regionen, so sagt die Vorhersagemaschine schon 33% aller Regentage falsch vorher. Kurz: Da wo die Regenvorhersage wichtig ist, sei es um einen Regenschirm mitzunehmen, sei es für die Landwirtschaft oder den Bau, ist das ansonsten „signifikant„ arbeitende Gerät recht ungenau. Und in nassen Regionen, wo es zu 90% immer regnet, ist dagegen der Anteil falscher Vorhersagen nur 0,7%. Daraus folgt: Wo es *meist* regnet, macht die Maschine wenig Fehler. Das bedeutet aber auch, dass man im Grunde auf sie verzichten kann, weil die Regenvorhersage pi mal Daumen eine eben so hohe Trefferquote hat - woraus weiter folgt: Ohne Vor-Urteil und Berücksichtigung von Rahmenbedingung machen „statistisch„ nachgewiesene Aussagen wenig Sinn.

Der Betrug fängst also schon dann an, wenn behauptet wird, eine Studie sei „vorurteilsfrei". Die Autoren zeigen, dass der Ergebniswert von Studien nur (!) dann beurteilt werden kann, wenn die Voraussetzungen oder Vor-Urteile und die Rahmenbedingungen angegeben sind.

Problematisch wird das insbesondere bei medizinischen Studien, z.B. bei den routinemäßigen Tests auf die Wirksamkeit neuer Arzneimittel. Das ist ein typischer Fall. Nimm 20 Patienten, gebe den einen das lange schon auf dem Markt erhältliche Mittel Altroruin und den anderen die Neuentwicklung Neufortuin. Vergleiche dann, wie viele von den Patienten mit dem einen oder anderen Mittel genesen, stecke die Zahlen in ein Statistikprogramm (keiner weiß, wie das eigentlich rechnet), es kommt $p < 5\%$ heraus, veröffentliche das Ergebnis und hole deinen Lorbeerkranz ab. Nicht untersucht werden hingegen die ganz entscheidenden Fragen, nämlich a) wie häufig tritt die untersuchte Krankheit in einer Region, einem Land oder weltweit überhaupt auf oder b) welche Aussagekraft dieses Ergebnis hat.

Dieses Dilemma der „schwachen" Studien (mit wenig „power„) ließe sich sicherlich über sehr viel größere Fallzahlen lösen, doch verbietet sich das meist aus finanziellen und manchmal aus ethischen Gründen.

Die Alternative, die die Autoren hier anbieten, kann allerdings auch nicht generell befriedigen. Sie setzt an dem altbekannten Schema von Erfolg und Irrtum (in Fachkreisen bekannt als „trial and error„) an. Wenn man mit einer Sache Erfolg hatte, bleibe man (statistisch überwiegend) dabei und wechsle nur bei Misserfolgen. („never-change-a-winning-team"). Sicherlich kann eine solche „pragmatische Optimierungsstrategie" besser und oft auch schneller bessere Resultate ergeben als die Schrotschuss-Strategie der statistischen Studien. Allerdings setzt auch diese von den Autoren vorgeschlagene Methode bestimmte Anwendungsbedingungen voraus, z.B. standardisierte Fälle, die wiederholt werden können.

Der Schein der Weisen, auf dem ganze Wissenschaftsgebilde beruhen, entpuppt sich somit als der „Scheiß zum Weinen" - wie die Autoren ganz drastisch formulieren. Die Wendung stammt aus einem der Dialoge verschiedener Personen, die sich durch das ganze Buch ziehen.

Wer allerdings bei den obigen Ausführungen nicht ganz mitgekommen ist, wird verstehen, wenn er das Buch liest. Die abstrakte und gemeinhin abschreckende Statistik machen die Autoren ganz konkret und lebendig - anhand und durch sehr anschauliche, oft auch lustige Geschichten. Da geht es um die schon erwähnte Regenvorhersage in trockenen und nassen Regionen eben so wie um die Überführung eines Mörders, das Angeln von Lecker-Fischen in Teichen, die voller Ekel-Fischen sind oder darum, wie hoch deren Überraschungswert („Informationsgehalt") ist, wenn ein Freund sich verspätet und nicht zum verein-

barten Termin eintrifft. So wird den LeserInnen das Problem mit der Statistik auf unterhaltsame Weise Schritt für Schritt näher gebracht bis sie plötzlich verstehen und die Tricks durchschauen lernen. Und das sollte man nicht verachten!

Vielleicht ist es die lockere Darstellungsweise, die dazu geführt hat, dass die umwälzenden Thesen der Autoren in der Fachwelt noch nicht richtig ernst genommen werden. Vielleicht aber auch aus dem sehr viel tiefer liegenden Grund, dass man dann alle diese Studien in den Papierkorb werfen und beginnen müsste, ernsthaft zu forschen - ober wie auch immer. Eines aber steht fest: Es ist gerade die sehr anschauliche und verständliche Form, die das Buch nutzbar macht als obligates Standardwerk in allen Studiengängen und Wissenschaftsbereichen mit statistischen Untersuchungen.

Noch gar nicht in letzter Konsequenz bedacht ist allerdings, wie viele Fehl-URTEILE von Gerichten aufgrund unzureichender angeblich „signifikanter" Studien gefällt wurden und werden, denn nicht nur im Berufskrankheitenrecht und der Arbeitsmedizin hat das „Signifikanzniveau" nahezu den Stellenwert einer göttlichen Größe. Tatsächlich sollte jeder Richter, jede Richterin in jedem Sozialgerichtsprozess daraufhin befragt werden, ob sie dieses Buch gelesen (und verstanden) haben, und wenn nicht, beantragen, die Sitzung auf den Tag nach erfolgter Lektüre zu vertagen.

Doch auch ohne wissenschaftlichen oder juristischen Hintergrund bietet das Buch gerade Lesern ohne mathematische Vorbildung (aber ohne Abneigung gegen die Mathematik) amüsanten Lesestoff. Viele Dinge, die man in Zeitungen liest, relativieren sich augenblicklich, wenn man nach dem wirklichen Informationsgehalt fragt. So würde z.B. die Wahrscheinlichkeit für die Vorhersage von Seitensprüngen von Bundeskanzlern sowohl von der allgemeinen Seitensprungwahrscheinlichkeit („Vor-Urteil„) wie auch von der Treffsicherheit von BILD-Zeitungsredakteuren für richtige Enthüllungen („power„) abhängen - danach fragt aber keiner.

So gesehen, bestimmt meist falsch verstandene Statistik unseren Alltag, Politik, Medien und Wissenschaft sehr viel mehr als wir denken. -

Redaktion Crash, Dr. Angela Vogel (verantw.), hrsg. von ***abeKra***, Verband arbeits- und berufsbedingt Erkrankter e.V.,
eMail:abekra-verband@t-online.de; http://www.abekra.de